```
Data Preprocessing -> Collecting & manipulating data -> Meaningful Information
Data Mining -> Exraction of pattern & knowledge from large amount of data.
Data preprocessing steps Converting a Raw Data -> Clean dataset
             from sklearn preprocessing import
Rescaling -> attributes with differnt scale are processed
Standardizing -- the dataset
                                       Standard Scaler Min Max Scaler Max Abs Scaler
                                       Robust Scaler (outliers)
Normalizing -> the dataset Normalizer
Encoading -> Categorical -> Integer One Hot Encoder
                                                              Lable Encoder
                                 Simple Imputer by defaut -> stratergy = mean
Handiling missing → data
                  Univariate 4
                                 Iterative Imputer ---- multivariate feature imputation
Removing duplicated -> data points
Removing Outliers -- handling noisy data
Discretize -> Data quantization or binning contineous data -> bins
               k Bins Discretizer
Split → Train-test split
                                   skearn · model_selection -> train_test_split
Steps in feature selection
Data Preprocessing ->
                         Feature scoring --- Selection
 Clean & prepare
                         Compute scores
                                                 most imp features based on
 data for feature sel<sup>n</sup>
                         reflets importance their scores
                          to target value
```

```
Feature Engg. -> create new features — in order to help ML models
       Filling missing values \rightarrow within variables
       Encoading categorical variables
       Variable transformation 📝
Feature Selection - from feature pool - predict target variables efficiently
Filter method based on statistical measures & correlations
 Information gain -> + redn in entropy
  from Sklearn feature_selection import mutual_info_classif
fisher score -> algorithm returns ranks of variables wirt. ____ decending order
 from sklearn feature _ selection import fisher _ score
 Chi-squared test -> used for categorical variables
      selection -> based on best chi-square score
  from sklearn feature _ selection import Select k Best
  from sklearn feature _ selection import chi2
 Correlation Coefficients -> df. corr() seln based on 0.5 threshold
      drop features having lower co-efficient value.
 Variance threshold → removes who doesn't meet some threshold.
       removes all zero-variance features.
       small drawback doesn't consider relationship b/w feature & target variable
         [mean absolute diffence]

Dispresion Ratio 

Geometric mean
     Similar to variance method
```

Feature engg. Vs Feature Selection

```
Wrapper method
```

```
Forward Feature Selection forward = True
```

iterative method -> starts with best performing features against target value.

Next selects another variable gives best performance combination.

### Backword Feature Elimination forward = False

iterative method -> storts with all available features against target value.

Next removes one variable which improves performance combination.

from mixtend. feature - selection import Sequential Feature Selection

#### Exhaustive feature selection

most robust feature seln method so far. -> Tries every possible combination returns best performing set.

from mlxtend feature\_selection import Exhaustive Feature Selection

#### Embedded Methods

## LASSO Regularization L1

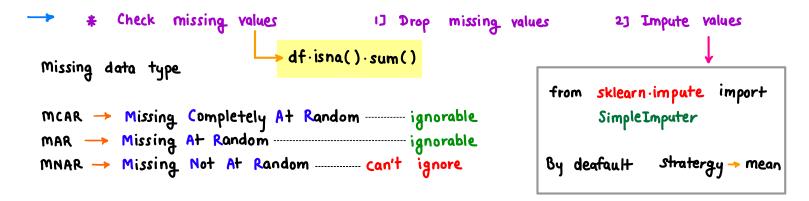
from sklearn · feature\_selection import SelectFromModel
model = SelectFromModel (model, prefit = True)

Selected\_columns = selected\_features.columns [selected\_features.var() != 0]

Random forest importance model -> rf

model · fit (x,y) -> importances = model · feature\_importances\_

# How we deal with missing data?



Pattern	Data Explains Pattern		
to missingness	yes	no	
<b>પુ</b> હ	MAR	MNAR	
No	-	MCAR	

Data Explains -> Other variables in dataset can explain pattern

example pollitical poll - MAR many people refuse to answer

# 1] Drop missing values

- 1> dropping rows where there are missing values -> df.dropna()
- 2> dropping entire row/column for multiple missing values in the row df.dropna (axis=0, thresh=2) drops rows where 🛨 >> 2
- 3> drop entire column

#### 2] Impute values

- 1> Replace NaN with given value
- df·col\_name·fillna(0) ← for numeric

df·col\_name·fillna ('string') ← for strings

- 2> Replace with mean, median, mode
  - df · col\_name · fillna ((df · col\_name · mean()))
- 3> Replace with previous or next value 'ffill' previous value Forward Fill df · col\_name · fillna (method = 'ffill') next value Backword Fill `bfill'
- 4> Interpolation df · col\_name · interpolate ( )

Scale	Example	Description	What is meaningful?	Central tendecy
Nominal	Name, Gender, Pin Code	Denotes name or gender. Ordering	only count	Mode
Ordinal	Low, medium , high Rank→ 1,2,3	value the orders  Ordering	only ordering can't measure dist. b/w values	Median Mode
Interval	Temperature °c	True zero absent	only diff. is meaningful but not the ratio	Mean Median Mode
Ratio Scale	Height , Weight Temperature kelvin	True zero present	Both diff. & ratio are meaningful	Mean Median Mode